A Gloss-driven Indian Sign Language Production System Using Learned Pose Representations

Suvajit Patra[1] , **Arkadip Maitra**[2] , **Swami Punyeshwarananda**[1] and **Soumitra Samanta**[1]
[1]Ramakrishna Mission Vivekananda Educational and Research Institute, Belur, India
[2]Red Hat
{suvajit.patra.cs20, punyeshwarananda, soumitra.samanta}@gm.rkmvu.ac.in, amaitra@redhat.com

Abstract

Sign Language Production (SLP) system translates spoken or written language into sign language, enabling accessible communication between the deaf/hard-of-hearing and the hearing population. Being one of the most widely used sign languages globally, Indian Sign Language (ISL) is a very low-resource language and lacks such SLP systems. This paper presents a scalable and modular SLP framework based on Sign-Pose-VQ-VAE model, designed for low-resource settings. The model learns discrete pose representations (codes) by disentangling body, left-hand, and right-hand keypoints, enabling efficient pose modeling and co-articulated sign generation. The proposed system is evaluated using a Hindi movie subtitle corpus coupled with an off-the-shelf back-translation model and achieves a gloss BLEU-4 score of 47.20. The system-generated signs are evaluated by certified ISL interpreters with an average rating of 4.33/5, and a BERT precision of 0.7683 on glosses. In addition, the proposed system achieves state-of-the-art performance among keypoint-based methods on the PHOENIX14T benchmark, attaining a BLEU-4 score of 10.03 and surpassing the previous best method by 0.67 points. The system with code is available at https://cs.rkmvu.ac.in/~isl/sl_gen_vqvae.

1 Introduction

Sign languages (SLs) serve as the primary means of communication for Deaf and hard-of-hearing communities. According to Ethnologue, Indo-Pakistani SL is the most widely used SL globally[1]. In India alone, an estimated 63 million individuals experience hearing impairments[2], yet access to inclusive education and public services remains severely limited. While SL interpretation has traditionally relied on human interpreters, their limited availability, high cost, and practical constraints highlight the urgent need for automatic SLP systems. By enabling real-time translation of spoken or written content into SL, SLP systems can democratize access to information and directly support the United Nations Sustainable Development Goals (SDGs)[3], particularly SDG 4 (Quality Education), SDG 8 (Decent Work), and SDG 10 (Reduced Inequalities) [Fink *et al.*, 2023]. This vision aligns with India's National Education Policy (NEP) 2020[4], which emphasizes accessibility and equity in learning and public service delivery.

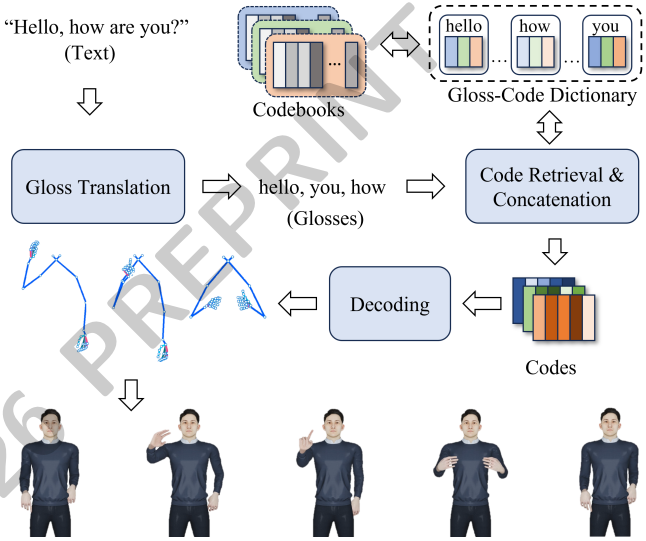

Figure 1: Overview of the proposed SLP system.

To the best of our knowledge, only a limited number of SLP systems exist for Indian Sign Language (ISL) [Sugandhi *et al.*, 2020; Kumar *et al.*, 2021]. These systems primarily rely on HamNoSys [Hanke, 2004], which requires expert linguistic annotation and thus limits scalability. Moreover, they typically generate utterances by concatenating word-level (isolated) signs without modeling co-articulation[5], resulting in temporally discontinuous and visually unnatural avatar animations. Recent deep learning based approaches [Stoll *et al.*, 2018; Zelinka and Kanis, 2020; Saunders *et al.*, 2020; Walsh *et al.*, 2024] have demonstrated promising results by

[1]https://www.ethnologue.com/guides/ethnologue200
[2]https://www.who.int/india/campaigns/world-hearing-day-2023
[3]https://sdgs.un.org/goals
[4]https://www.education.gov.in/en/nep/about-nep
[5]Smooth transitions between isolated signs.

leveraging large, well-annotated datasets such as RWTH-PHOENIXWeather-2014T (PHOENIX14T) [Camgoz *et al.*, 2018] and others [Duarte *et al.*, 2021; Zhou *et al.*, 2021; Albanie *et al.*, 2021]. However, comparable large-scale, well-annotated continuous datasets are unavailable for ISL. As a result, ISL remains a severely low-resource language for SLP. Additionally, many existing systems require retraining to incorporate vocabulary expansion, further limiting scalability and real-world deployment.

Based on the output modality, SLP methods can be broadly categorized into three groups: (1) **video-based** (generate RGB frame sequences) [Saunders *et al.*, 2022; Shi *et al.*, 2024](2) **keypoint-based** (generate sequences of human pose keypoints) [Zelinka and Kanis, 2020; Saunders *et al.*, 2020; Xie *et al.*, 2024; Shi *et al.*, 2024; Tang *et al.*, 2025c; Wang *et al.*, 2025; Tang *et al.*, 2025b]; and (3) **avatar-based** (produce animations of digital 3D humanoid avatars) [Zuo *et al.*, 2024; Yin *et al.*, 2024; Sugandhi *et al.*, 2020]. Across paradigms, generating temporally coherent pose keypoint sequences from glosses[6] remains a core challenge, as pose generation serves as a common intermediate step.

Sign language is inherently visual and relies not only on isolated signs but also on smooth transitions between them (co-articulation). Proper co-articulation is therefore essential for synthesizing meaningful continuous signs [Tang *et al.*, 2025a]. In low-resource ISL settings, well-annotated continuous datasets are scarce, whereas several isolated sign datasets [Sridhar *et al.*, 2020; Joshi *et al.*, 2022; Patra *et al.*, 2025b] are publicly available. In this work, we leverage these isolated resources to construct sentence-level signs.

Gloss-free end-to-end models [Saunders *et al.*, 2020; Tang *et al.*, 2025b] that directly regress pose sequences often suffer from temporal inconsistency and regression-to-the-mean effects. Discretizing pose space using Vector Quantized Variational Autoencoder (VQ-VAE) [Van Den Oord *et al.*, 2017] alleviates these issues by learning discrete pose tokens [Rastgoo *et al.*, 2024]. However, subtle hand and finger motions, critical for SL, are frequently attenuated due to their high-frequency and low-amplitude nature [Zhang *et al.*, 2025]. To address this, we disentangle body, left-hand, and right-hand keypoints and normalize hand poses to better capture handshape and fine-grained finger articulation.

Recent dictionary-based approaches [Walsh *et al.*, 2024; Zuo *et al.*, 2024] demonstrate that stitching isolated signs avoids regression-to-the-mean issues and enables scalable vocabulary expansion. Inspired by this paradigm, a dictionary-based SLP system is proposed that produces natural continuous signing with proper co-articulation. This work makes the following key contributions:

- A modular and scalable SLP system for low-resource sign languages, demonstrated on Indian Sign Language, that generates sentence-level continuous signs without requiring large continuous datasets.
- Sign-Pose-VQ-VAE: a VQ-VAE-based model that learns discrete body, left-hand, and right-hand pose codebooks, enabling fine-grained hand articulation and natural co-articulation.
- A simple baseline method to transfer pose trajectories to 3D avatar animation through bone-angle replication for enhanced end-user understanding.
- Comprehensive evaluation demonstrating strong performance, including a gloss BLEU-4 of 47.20 on ISL, ISLRTC[7] certified interpreter-rated quality of 4.33/5, and state-of-the-art results among keypoint-based approaches on PHOENIX14T (BLEU-4 score of 10.03).

This project is part of a larger initiative to enable bidirectional translation between spoken language and ISL in educational settings, initiated by the ISL unit of FDMSE, RKMVERI, Coimbatore, a teaching institution for disability management in India. In addition to ISL domain experts, certified ISL interpreters (ISLRTC-certified) and professional linguistic experts contributed by annotating gloss translations from English sentences and evaluating the system.

The rest of the paper is organised as follows: Section 2 discusses the related works, and Section 3 describes the proposed method. Experimental evaluation is presented in Section 4, and the paper concludes in Section 5.

2 Related Works

Prior SLP methods [Hu *et al.*, 2022; Cox *et al.*, 2002; Glauert *et al.*, 2006] which rely on motion capture data requires specialized, expensive equipments, restricting implementations in low-resource settings. Other methods [Zwitserlood *et al.*, 2004; Goyal and Goyal, 2016; Sugandhi *et al.*, 2020; Kumar *et al.*, 2021] use the Hamburg Notation System (HamNoSys) [Hanke, 2004], which often produces robotic signing with poor co-articulation, resulting in unnatural output that is generally not well received by the deaf community.

With deep learning advances, researchers are exploring its potential in SLP. The neural machine translation based approach [Stoll *et al.*, 2018] splits the task into Text-to-Gloss (T2G) using a Recurrent Neural Network, Gloss-to-Pose (G2P) using a gloss-2D pose dictionary, and Pose-to-Sign using a Generative Adversarial Network (GAN) [Goodfellow *et al.*, 2014]. However, it ignores hand keypoints, crucial in SL. Progressive Transformers [Saunders *et al.*, 2020] uses two Transformers [Vaswani *et al.*, 2017] for T2G and G2P. Tang *et al.*'s [Tang *et al.*, 2022] GEN-OBT enhances a transformer gloss encoder and applies a reverse-translation objective to better align gloss and pose representations. Viegas *et al.* [Viegas *et al.*, 2023] incorporate facial expressions as non-manual features in a transformer-based SLP model.

Recent works in SLP have moved rapidly from gloss-conditioned encoder–decoder pipelines toward diffusion-based generative models. Xie *et al.* [Xie *et al.*, 2024] used Pose-VQVAE and a discrete denoising diffusion model (G2P-DDM), using vector-quantized codes and a CodeUNet to capture spatial–temporal structure while handling variable output length. GCDM [Tang *et al.*, 2025c], and Sign-IDD [Tang *et al.*, 2025b], are other notable works that show the effectiveness of diffusion mechanisms for the SLP task. Despite their promising results, these models often generate temporally inconsistent pose sequences. In contrast, we select a convolu-

[6] A written label representing a unit sign for annotation.

[7] https://islrtc.nic.in/directory-of-isl-interpreters/

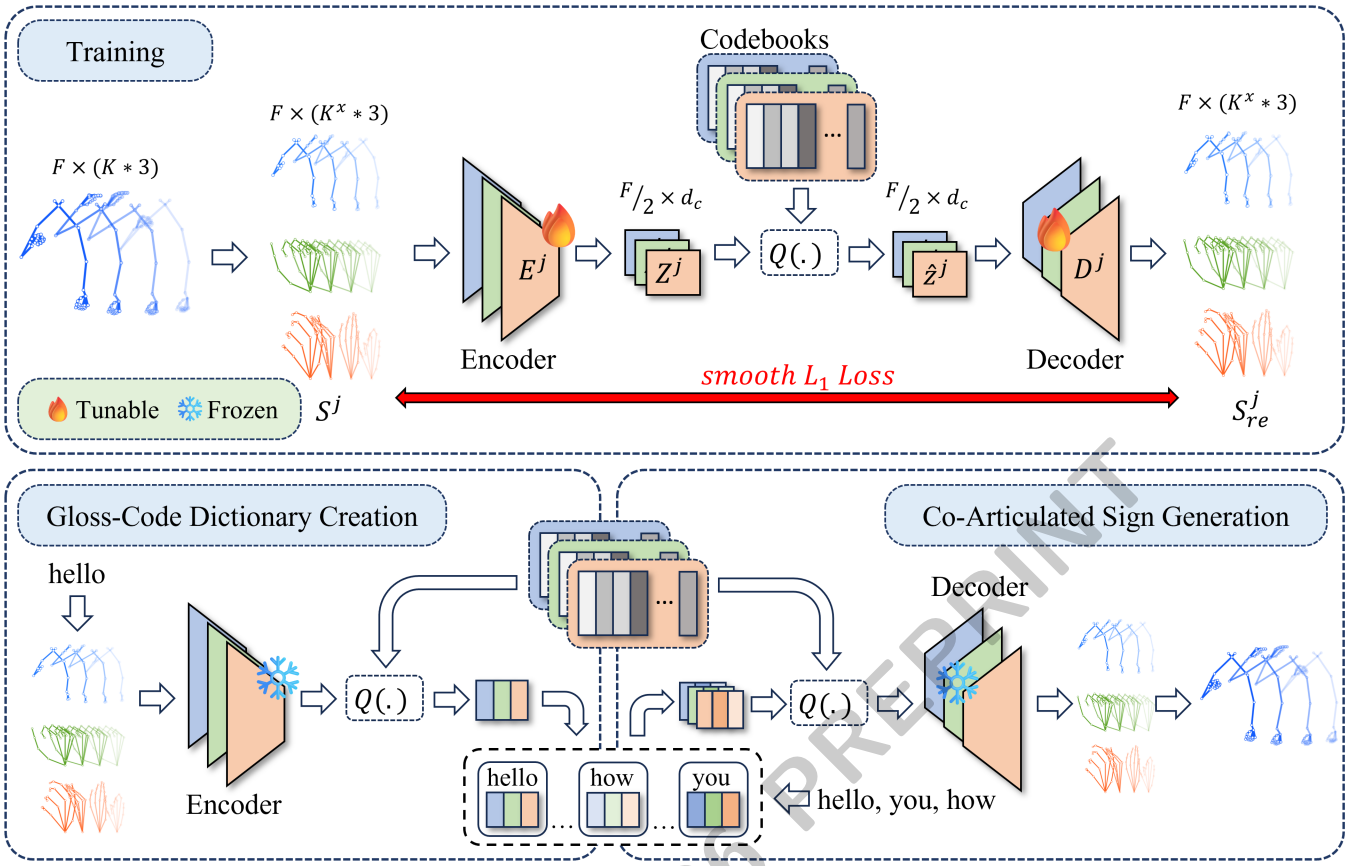

Figure 2: Overview of Sign-Pose-VQ-VAE training, gloss-code dictionary creation, and co-articulated sign generation. During training, pose sequences are randomly segmented, split into body, left-hand, and right-hand streams, and processed by the corresponding VQ-VAE, with $K^x = 15, 21$, and 21 keypoints, respectively. For dictionary creation, isolated-sign pose sequences are encoded and quantized to obtain discrete codebook indices stored per gloss. During co-articulated sign generation, index sequences for a gloss sentence are retrieved, temporally concatenated, decoded, and combined to produce the final continuous sign pose sequence.

tional architecture along the temporal dimension for the Sign-Pose-VQ-VAE, which effectively captures local temporal dependencies and ensures smoother, more temporally consistent pose transitions throughout the generated sign sequences.

VQ-VAE [Van Den Oord *et al.*, 2017] based approach T2S-GPT [Yin *et al.*, 2024] uses a dynamic VQ-VAE to learn variable-length codes and use GPT networks to predict codes and lengths, from which they generate the continuous sign sequence. Here, they considered body keypoints, including hands, whereas we decoupled the keypoints into multiple streams, which leads to a wide variety of encoded poses in a relatively small codebook. Recently, Zuo *et. al.* [Zuo *et al.*, 2024] proposed a 3D avatar based approach. They segmented the continuous SL videos and created a gloss-3D avatar sign dictionary. They concatenate these 3D avatar signs using linear interpolation method. Similarly, in ISL, Patra et al. [Patra *et al.*, 2025a] leveraged an isolated keypoint-based sign dictionary and linear interpolation for co-articulation to generate continuous sign sequences. In contrast to these dictionary-based works, our approach only stores the codebook indices (tokens) in the dictionary, which is much more efficient in real-world applications.

3 Proposed Methodology

The proposed SLP system is formulated as follows: given a spoken-language sentence $X = [x_1, x_2, \ldots, x_W]$ consisting of W words, the goal is to generate an output sequence $Y = [y_1, y_2, \ldots, y_O]$ of length O, where each y_i represents a SL output unit, such as a skeleton pose, a photo-realistic frame, or, in our case, an avatar animation frame. The proposed system consists of *four* stages: gloss translation, gloss-code dictionary generation, co-articulated sign generation, and avatar animation generation. An overview of the proposed system is shown in Figure 1, where English text is translated into a sequence of glosses. The corresponding codes are then retrieved, concatenated, and decoded to generate a co-articulated sign sequence, which is finally converted into a signing avatar animation.

3.1 Gloss Translation

The goal of this stage is to translate spoken-language text into a sequence of sign glosses. Due to the lack of a well-annotated text-gloss parallel corpus for ISL, neural machine translation approaches are not directly applicable. Therefore, a rule-based method is adopted, and a large language model

(LLM)-based approach is additionally explored.

Rule-based: Each spoken-language sentence is parsed into a syntactic tree using the Stanza library[8] with part-of-speech tagging. The sentence is then reordered into a subject-object-verb (SOV) structure, followed by the application of ISL grammatical rules inspired by Sugandhi *et al.* [Sugandhi *et al.*, 2020] to generate ISL gloss sequences.

LLM-based: Inspired by Inan *et al.* [Inan *et al.*, 2024], an LLM-based approach for text-to-ISL-gloss translation is employed. With the assistance of ISL interpreters, a small corpus of 300 English sentences with their corresponding ISL gloss translations is manually created. The LLM is then prompted to translate unseen sentences into ISL glosses based on these examples.

3.2 Gloss-Code Dictionary Creation

Dictionary-based approaches [Saunders *et al.*, 2022; Walsh *et al.*, 2024; Zuo *et al.*, 2024; Inan *et al.*, 2024] typically store high-dimensional visual representations (e.g., frames, videos, animations, or pose sequences), which are memory-intensive and limit scalability. Instead, **Sign-Pose-VQ-VAE** is proposed, which learns compact, discrete representations of body poses and handshapes by disentangling body, left-hand, and right-hand keypoints into separate codebooks. Only the resulting codebook indices (tokens) are stored in a gloss-code dictionary, while continuous pose sequences with smooth co-articulation are reconstructed by a convolutional decoder. An overview of this stage is illustrated in Figure 2.

Sign Pose Vector Quantized Variational Autoencoder (Sign-Pose-VQ-VAE): Given a pose sequence $S = [s_1, \ldots, s_F]$ extracted from a sign video, where $s_i \in \mathbb{R}^{K\times 3}$ denotes K 3D keypoints, S is decoupled into *three* streams: body, left hand, and right hand. To preserve hand-body coordination, wrist joints and selected 2 metacarpophalangeal (MCP) joints from each hand are included in the body stream, resulting in 15 body keypoints. These *three* additional keypoints per hand are retained to later reattach the hands to the body. Hand keypoints are normalized per frame to achieve invariance to position, rotation, and scale.

The resulting streams are $S^b \in \mathbb{R}^{F\times 15\times 3}$, $S^l \in \mathbb{R}^{F\times 21\times 3}$, and $S^r \in \mathbb{R}^{F\times 21\times 3}$. Each stream is processed by an independent VQ-VAE with encoder E^j and decoder D^j ($j \in \{b, l, r\}$). The encoder consists of *three* 1D convolutions followed by a residual block [He *et al.*, 2016] and ReLU activations, producing latent features $Z^j = [z_1^j, \ldots, z_{F/2}^j]$, where $z_i^j \in \mathbb{R}^{d_c}$. Latents are quantized using a learnable codebook $C^j = \{c_n^j\}_{n=1}^N$:

$$\hat{z}_i^j = \arg\min_{c_n^j \in C^j} \|z_i^j - c_n^j\|_2. \tag{1}$$

Codebooks are updated using exponential moving average (EMA) [Razavi *et al.*, 2019] with codebook reset [Williams *et al.*, 2020] to prevent collapse. The decoder D^j reconstructs pose sequences using residual blocks, nearest-neighbor upsampling, *four* 1D convolution layers, and ReLU activations. Reconstruction is optimized with a smooth L_1 loss:

$$\mathcal{L}_{re}^j = \mathcal{L}_1^{smooth}(S^j, D^j(\hat{Z}^j)). \tag{2}$$

Since the EMA variant is used, the embedding loss is omitted, while the commitment loss is retained:

$$\mathcal{L}_{vq}^j = \mathcal{L}_{re}^j + \beta \underbrace{\|Z^j - sg[\hat{Z}^j]\|^2}_{\mathcal{L}_{commit}}, \tag{3}$$

where $sg[\cdot]$ denotes the stop-gradient operator and β is the weighting coefficient for the commitment loss.

After combining all three streams, the final objective function of Sign-Pose-VQ-VAE is:

$$\mathcal{L}_{vq} = \mathcal{L}_{vq}^b + \mathcal{L}_{vq}^l + \mathcal{L}_{vq}^r. \tag{4}$$

The proposed Sign-Pose-VQ-VAE is trained on randomly sampled pose segments from the isolated ISL dataset FDMSE-ISL [Patra *et al.*, 2025b]. After training, each isolated gloss v is encoded into *three* discrete index sequences (I^b, I^l, I^r), where $I^j = [i_1^j, \ldots, i_{F/2}^j]$ and $i_m^j \in \{1, \ldots, N\}$. The resulting gloss-code dictionary maps each gloss to its corresponding discrete pose tokens (Figure 2).

3.3 Co-Articulated Sign Generation

During training, the Sign-Pose-VQ-VAE decoders are trained to learn smooth pose and handshape transitions; when applied jointly, they implicitly model co-articulation across consecutive signs. An overview of this module is shown in Figure 2. Given a gloss sequence $V = [v_1, \ldots, v_G]$, the corresponding discrete code sequences are retrieved from the gloss-code dictionary. For each gloss v_i, *three* code sequences $(\hat{Z}_i^b, \hat{Z}_i^l, \hat{Z}_i^r)$ are obtained. These sequences are concatenated independently for the body and both hands, $\hat{Z}' = [||_{i=1}^G \hat{Z}_i^b, ||_{i=1}^G \hat{Z}_i^l, ||_{i=1}^G \hat{Z}_i^r]$, where $||$ denotes temporal concatenation. Each stream of $\hat{Z}'$ is decoded by its corresponding VQ-VAE decoder to generate continuous body pose and handshape sequences. Finally, the handshape streams are coupled with the body pose stream by aligning the shared keypoints, yielding a temporally smooth, co-articulated sign sequence.

3.4 Avatar Animation Generation

Given a continuous pose sequence, the goal is to generate a corresponding avatar animation. Since fine-grained hand articulation is critical for SL, the upper body, particularly the arms and hands, is animated. A generic humanoid avatar from Mixamo[9] is utilized, and human joint rotation constraints (e.g., *three* degrees of freedom for shoulders and one for elbows) are enforced to compute per-bone Euler angles. These angles are then applied to the avatar skeleton to produce the final animation.

4 Experiments & Results

4.1 Dataset

The proposed system is evaluated on two datasets: the FDMSE-ISL dataset [Patra *et al.*, 2025b] and the RWTH-PHOENIXWeather-2014T (PHOENIX14T) dataset [Camgoz

[8]https://stanfordnlp.github.io/stanza/

[9]https://www.mixamo.com

Method	BERT-P	BERT-R	BERT-F1	Time (s)
Rule-based	0.8575	0.8519	0.8546	**0.09**
GPT-4o-mini	**0.8768**	**0.8782**	**0.8774**	1.04
Qwen3-4B	0.8731	0.8746	0.8737	0.19

Table 1: Gloss translation evaluation results of different methods.

Separated Codebook			
Codebook Size	**Code Dimension**	**WER↓**	**BLEU-1↑**
512	128	34.36	64.27
512	256	32.62	66.19
512	512	35.02	63.47
512	786	36.63	62.11
786	128	35.49	63.16
786	256	31.19	67.80
786	512	32.63	66.51
786	786	32.76	66.13
1024	128	32.86	66.08
1024	256	31.49	67.42
1024	512	33.02	66.02
1024	786	32.44	66.41
1536	**128**	**27.90**	**71.49**
1536	256	29.77	69.72
1536	512	31.21	67.96
1536	786	29.04	70.15
Shared Codebook			
Codebook Size	**Code Dimension**	**WER ↓**	**BLEU-1 ↑**
1536	128	44.47	52.58
1536	384	44.34	53.13
4608	128	36.11	62.51

Table 2: Ablation study of Sign-Pose-VQ-VAE hyperparameters (codebook size and code dimension) for sign language recognition (pose→gloss) on the SynICS-20k dataset

et al., 2018]. The FDMSE-ISL dataset [Patra *et al.*, 2025b] contains 40,033 isolated ISL videos spanning 2,002 unique glosses. This dataset is used to train Sign-Pose-VQ-VAE and to construct the gloss-code dictionary. To evaluate sentence-level sign generation, a Synthetic Indian Continuous Signs dataset (SynICS-20k) is constructed by concatenating isolated signs corresponding to 20,000 generic English subtitle sentences derived from Hindi movies from OpenSubtitles[10]. Linear interpolation is applied at sign boundaries. This dataset is referred to as SynICS-20k and is split into training (70%), validation (10%), and test (20%) sets.

The RWTH-PHOENIXWeather-2014T (PHOENIX14T) dataset [Camgoz *et al.*, 2018] consists of 8,257 German sentences with corresponding gloss annotations and sign videos over 1,115 unique glosses, and is used for state-of-the-art benchmarking. This dataset has no isolated gloss-level video annotations and is segmented using the CTC-based force alignment method proposed by Zuo *et al.* [Zuo *et al.*, 2024].

4.2 Experiment Settings

Sign-Pose-VQ-VAE is implemented in PyTorch and trained using the AdamW optimiser with an initial learning rate of 2×10^{-4}, cosine annealing, 20 warm-up epochs, early stopping, and commitment loss coefficient $\beta = 0.02$. The model is trained for 7.5k epochs on SynICS-20k and 30k epochs on PHOENIX14T. Each stream is configured with a codebook of size $N = 1536$ and latent dimensionality $d_c = 128$, selected via ablation. All experiments are conducted on an Intel i9-13900K CPU with an NVIDIA RTX 4090 GPU. MediaPipe Holistic[11] is used for 3D keypoint estimation on the FDMSE-ISL dataset. For fair comparison on the PHOENIX14T dataset, HRNet [Sun *et al.*, 2019] is employed for 2D pose estimation, and a Recurrent Neural Network (RNN)-based pose correction model [Zelinka and Kanis, 2020] is used to convert to 3D keypoints.

4.3 Data Processing

For each sign video, pose sequences are extracted and 51 upper-body keypoints are retained: 21 per hand, 2 eyes, 1 nose, 2 shoulders, 2 forearms, and 2 wrists. Missing hand detections can introduce redundant codes in Sign-Pose-VQ-VAE; this issue is mitigated by filling missing keypoints using spherical linear interpolation (Slerp). Keypoints are normalised using the left shoulder (k_{ls}), right shoulder (k_{rs}), and nose (k_n) as $\{k_i - k_n - 3 \times norm(k_{ls}, k_{rs})\}/\{6 \times norm(k_{ls}, k_{rs})\}$, where k_i denotes any keypoint.

4.4 Evaluation Metrics

Following prior SLP work [Wang *et al.*, 2025; Tang *et al.*, 2025b; Tang *et al.*, 2025a; Saunders *et al.*, 2020; Walsh *et al.*, 2025], back-translation quality is evaluated using BLEU, WER, and ROUGE-L scores. For direct pose-level evaluation, we adapted standard metrics for sign language production: Dynamic Time Warping (DTW) handles temporal misalignment by optimally warping sequences, making it suitable for comparing pose motion performed at varying speeds, and Total Distance (TD) [Walsh *et al.*, 2025] measures the ratio of overall distance the signer's dominant hand has moved in 3D space in reference and hypothesis to judge how expressive the productions are. Additionally, following Inan *et al.* [Inan *et al.*, 2024], semantic similarity is measured using BERT-based precision, recall, and F1 scores (BERT-P/R/F1) [Zhang *et al.*, 2019] for both back-translation and gloss translation.

4.5 Ablation Study

For the text-to-gloss translation module, *three* strategies are explored: a rule-based method and two LLM-based models, GPT-4o-mini[12] (proprietary) and Qwen3-4B[13] (open-source). All methods are applied to the SynICS-20k test set, and semantic similarity between input sentences and generated glosses is measured using BERT scores (Table 1). GPT-4o-mini achieves the best performance, but its gains over the rule-based approach are modest (0.0193, 0.0263, and 0.0228 in BERT-P/R/F1). Qwen3-4B shows similar marginal improvements. In contrast, the rule-based method is substantially more efficient, being approximately 12× faster than

[10]https://opus.nlpl.eu/OpenSubtitles/hi&en/v2024/OpenSubtitles

[11]https://ai.google.dev/edge/mediapipe/solutions/vision/holistic_landmarker

[12]https://platform.openai.com/docs/models/gpt-4o-mini

[13]https://github.com/QwenLM/Qwen3

	Syntactic				Semantic			Direct Pose	
Method	**WER↓**	**BLEU-1↑**	**BLEU-4↑**	**ROUGE-L↑**	**BERT-P↑**	**BERT-R↑**	**BERT-F1↑**	**DTW↓**	**TD↑**
Linear	25.27	75.03	50.66	79.08	0.9381	0.9255	0.9316	0.000	0.000
PT [Saunders *et al.*, 2020]	59.21	34.83	7.28	49.92	0.8785	0.8430	0.8601	1.118	0.372
Sign-IDD [Tang *et al.*, 2025b]	36.62	62.44	35.36	69.65	0.9176	0.8940	0.9054	0.966	0.579
Sign-Pose-VQ-VAE (Ours)	**27.90**	**71.49**	**47.20**	**77.95**	**0.9373**	**0.9186**	**0.9276**	**0.943**	**0.662**

Table 3: Sign language recognition (pose→gloss) and direct pose comparison on SynICS-20k test set. Syntactic and Semantic metrics are calculated on generated gloss sequences. ↑/↓ indicates higher/lower is better.

	Dev				Test			
Method	**WER↓**	**BLEU-1↑**	**BLEU-4↑**	**ROGUE-L↑**	**WER↓**	**BLEU-1↑**	**BLEU-4↑**	**ROGUE-L↑**
Ground Truth	61.92	34.93	13.10	35.87	60.53	34.35	12.44	34.79
PT-base [Saunders *et al.*, 2020]	98.53	9.53	0.72	8.61	98.36	9.47	0.59	8.88
PT-GN [Saunders *et al.*, 2020]	96.85	12.51	3.88	11.87	96.50	13.35	4.31	13.17
NAT-AT [Huang *et al.*, 2021]	-	-	-	-	88.15	14.26	5.53	18.72
NAT-EA [Huang *et al.*, 2021]	-	-	-	-	82.01	15.12	6.66	19.43
D3DP-sign [Shan *et al.*, 2023]	91.51	17.20	5.01	17.94	91.83	16.51	5.25	17.55
DET [Viegas *et al.*, 2023]	-	-	-	-	-	17.18	5.76	17.64
G2P-DDM [Xie *et al.*, 2024]	-	-	-	-	77.26	16.11	7.50	-
GDCM [Tang *et al.*, 2025c]	82.81	22.88	7.64	23.35	81.94	22.03	7.91	23.20
GEN-OBT [Tang *et al.*, 2022]	82.36	24.92	8.68	25.21	81.78	23.08	8.01	23.49
Sign-IDD [Tang *et al.*, 2025b]	77.72	25.40	8.93	27.60	76.66	24.80	9.08	26.58
LVMCN [Wang *et al.*, 2025]	76.86	25.79	9.17	27.29	75.43	24.33	9.36	26.24
Sign-Pose-VQ-VAE (Ours)	**71.47**	**32.69**	**10.87**	**32.72**	**72.15**	**31.45**	**10.03**	**31.20**

Table 4: Comparison of back-translation (pose→gloss→text) performance on PHOENIX14T. Metrics are computed on generated gloss sequences (WER) and generated text (BLEU-1, BLEU-4, and ROUGE-L). ↑/↓ indicates higher/lower is better.

GPT-4o-mini and 2× faster than Qwen3-4B. Given its competitive performance and significantly lower computational cost, the rule-based approach is adopted for all subsequent experiments.

An extensive ablation study is conducted on the two main hyperparameters of Sign-Pose-VQ-VAE: codebook size and code dimension. Pose-to-Gloss translation is evaluated using the Sign Language Transformer [Camgoz *et al.*, 2020]. Results are reported in Table 2. We compare a shared codebook (single stream across all keypoints) with separated codebooks (independent streams for body and hands). Separated codebooks consistently outperform the shared configuration across both WER and BLEU-1. The best performance is achieved with a codebook size of 1536 and a code dimension of 128, which is used in all experiments.

4.6 Quantitative Evaluation

The proposed method is compared against Progressive Transformers (PT) [Saunders *et al.*, 2020] and the diffusion-based Sign-IDD [Tang *et al.*, 2025b]. Both methods are trained on SynICS-20k with author-suggested optimal hyperparameter settings. Quantitative evaluation is performed via back-translation using Sign Language Transformers (SLT) [Camgoz *et al.*, 2020], trained on the synthetic dataset and evaluated on the linear synthetic test set, as well as on test sets generated by PT, Sign-IDD, and the proposed method. Results are summarised in Table 3. Across all metrics, the proposed approach significantly outperforms both baselines. Compared with PT, absolute improvements of 31.31 WER, 39.92 BLEU-4, and 0.0588 BERT-P are achieved. Relative to Sign-IDD, gains of 8.72 WER, 11.84 BLEU-4, and 0.0197 BERT-P are obtained. Direct pose-level evaluation using DTW and TD on the dominant signing hand is shown in Table 3. The proposed method achieves the lowest DTW (0.94) and best TD (0.66), indicating superior temporal alignment and more expressive hand motion.

Additionally, the system is evaluated on the PHOENIX14T dataset using SLT back-translation, and results are reported in Table 4. Compared to the transformer-based PT [Saunders *et al.*, 2020], the proposed method achieves 27.06 points WER reduction and 10.15 points BLEU-4 improvement on the validation set, and 26.21 points WER reduction and 9.44 points BLEU-4 improvement on the test set. Against graph-based methods (NAT-AT [Huang *et al.*, 2021], NAT-EA [Huang *et al.*, 2021]), the proposed method outperforms the strongest baseline, NAT-EA, by 9.86 points lower WER and 3.37 points higher BLEU-4 on the test set. Compared to diffusion-based methods (D3DP-sign [Shan *et al.*, 2023], G2P-DDM [Xie *et al.*, 2024], GDCM [Tang *et al.*, 2025c], Sign-IDD [Tang *et al.*, 2025b]), the proposed approach surpasses the best model, Sign-IDD, with 6.25 points WER reduction and 1.94 points BLEU-4 improvement on the validation set, and 4.51 points WER reduction and 0.95 points BLEU-4 improvement on the test set. Finally, compared to the state-of-the-art method LVMCN [Wang *et al.*, 2025], the proposed method delivers 5.39 points lower WER and 1.70 points higher BLEU-4 on the validation set, and 3.28 points lower WER and 0.67 points higher BLEU-4 on the test set.

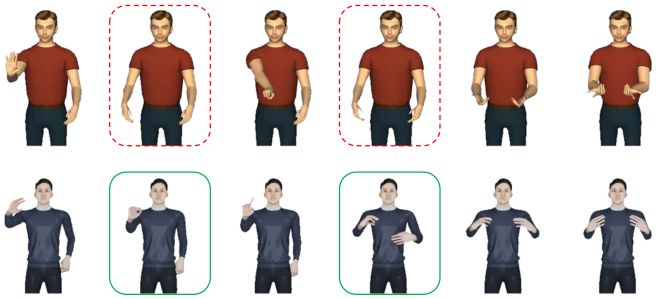

Figure 3: Visual comparison of avatar-based sign generation for Indian sign language on the sentence "Hello, how are you?". The top animation, produced by the previous HamNoSys-based method returns to a rest pose between signs (red dashed box), producing discontinuous signing, while the bottom animation, generated by the proposed system, generates smooth co-articulated transitions between signs (green box), resulting in continuous signing

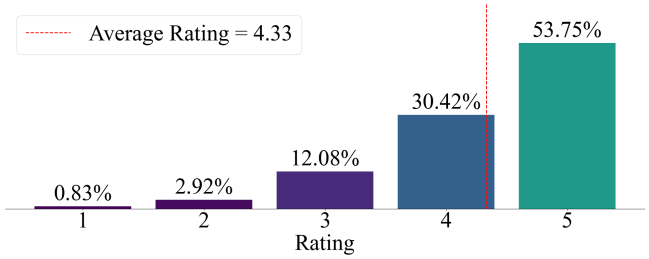

Figure 4: Distribution of experts' ratings of the proposed SLP system.

4.7 Qualitative Evaluation

The acceptability of the proposed SLP system in the Deaf community is assessed with *three* ISLRTC[14] certified ISL experts using 240 English sentences from a Hindi movie subtitles corpus. A web-based application was developed for this evaluation, where for each sample, the input text is presented alongside the generated avatar animation and a stick-figure visualization. Experts assign a score based on the rubrics: 1 - Not interpretable/minimally interpretable; 2 - Poor (some isolated words interpretable but overall intent unclear); 3 - Modest (overall intent clear but noticeable gestural errors and/or sign ordering error); 4 - Good (overall intent clear but some minor gestural errors); 5 - Excellent (fully interpretable with all glosses distinguishable). For samples rated below 5, experts additionally provide the glosses they interpreted from the animation, along with the expected glosses corresponding to the input text. The experts report an average rating of 4.33, with the rating distribution shown in Figure 4. Semantic similarity between the interpreted glosses (hypotheses) and expected glosses (references) is computed using BERT score, yielding BERT-P = 0.7683, BERT-R = 0.7497, and BERT-F1 = 0.7585.

Figure 3 compares the proposed approach with the HamNoSys-based avatar animation [Sugandhi *et al.*, 2020] in ISL. The prior method resets the avatar to a rest pose after each sign (e.g., the first row, columns 2 and 4 in Figure 3), resulting in disjointed motion. In contrast, the proposed system generates temporally smooth transitions with effective co-articulation, resulting in more natural and continuous signing.

[14]https://islrtc.nic.in/directory-of-isl-interpreters/

5 Conclusion

This work presents a scalable SLP system for low-resource settings that translates spoken or written language into 3D sign language avatar animations using isolated signs. The proposed Sign-Pose-VQ-VAE model learns discrete body poses and hand shapes while enabling smooth, co-articulated continuous sign generation. The system achieves state-of-the-art performance on standard translation metrics for both Indian Sign Language (ISL) and German sign language. Overall, this approach overcomes key limitations of prior methods, particularly in fine-grained hand articulation, co-articulation quality, and scalability. The system is evaluated by ISLRTC certified ISL experts, receiving high ratings and positive feedback regarding its societal impact. The system with code is publicly available at https://cs.rkmvu.ac.in/~isl/sl_gen_vqvae, aiming to strengthen accessibility and inclusion for the Deaf community, aligning with global goals for equitable communication.

Ethical Statement

This work uses the FDMSE-ISL [Patra *et al.*, 2025b] and PHOENIX14T [Camgoz *et al.*, 2018] datasets, which are publicly available. According to the authors, these datasets were collected with informed consent from the signers. We ensure participant confidentiality by using only pose sequences and avatar animations, without sharing any identifiable visual data.

Acknowledgments

We would like to thank Ms. Indira Ghosh for her assistance in creating the English sentence to ISL gloss corpus. We gratefully acknowledge the support from the Department of Science and Technology (DST), Government of India, through the DST-FIST grant [Sanction No.: SR/FST/MS-I/2022/116].

References

[Albanie *et al.*, 2021] Samuel Albanie, Gül Varol, Liliane Momeni, Hannah Bull, Triantafyllos Afouras, Himel Chowdhury, Neil Fox, Bencie Woll, Rob Cooper, Andrew McParland, and Andrew Zisserman. BOBSL: BBC-Oxford British Sign Language Dataset. *arXiv*, 2021.

[Camgoz *et al.*, 2018] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

[Camgoz *et al.*, 2020] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2020.

[Cox *et al.*, 2002] Stephen Cox, Michael Lincoln, Judy Tryggvason, Melanie Nakisa, Mark Wells, Marcus Tutt, and Sanja Abbott. Tessa, a system to aid communication with deaf people. In *Proceedings of the International ACM Conference on Assistive Technologies*, pages 205–212, 2002.

[Duarte *et al.*, 2021] Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. How2Sign: A Large-scale Multimodal Dataset for Continuous American Sign Language. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2021.

[Fink *et al.*, 2023] Jerome Fink, Pierre Poitier, Maxime André, Loup Meurice, Benoît Frénay, Anthony Cleve, Bruno Dumas, and Laurence Meurant. Sign language-to-text dictionary with lightweight transformer models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 5968–5976, 2023.

[Glauert *et al.*, 2006] John RW Glauert, Ralph Elliott, Stephen J Cox, Judy Tryggvason, and Mary Sheard. Vanessa–a system for communication between deaf and hearing people. *Technology and Disability*, 18(4):207–216, 2006.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014.

[Goyal and Goyal, 2016] Lalit Goyal and Vishal Goyal. Automatic translation of English text to Indian Sign Language synthetic animations. In *Proceedings of the International Conference on Natural Language Processing*, pages 144–153, 2016.

[Hanke, 2004] Thomas Hanke. Hamnosys–representing sign language data in language resources and language processing contexts. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 1–6, 2004.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[Hu *et al.*, 2022] Li Hu, Jiahui Li, Jiashuo Zhang, Qi Wang, Bang Zhang, and Ping Tan. A speech-driven sign language avatar animation system for hearing impaired applications. In *IJCAI*, pages 5912–5915, 2022.

[Huang *et al.*, 2021] Wencan Huang, Wenwen Pan, Zhou Zhao, and Qi Tian. Towards fast and high-quality sign language production. In *Proceedings of the ACM International Conference on Multimedia*, page 3172–3181, 2021.

[Inan *et al.*, 2024] Mert Inan, Katherine Atwell, Anthony Sicilia, Lorna Quandt, and Malihe Alikhani. Generating signed language instructions in large-scale dialogue systems. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 140–154, 2024.

[Joshi *et al.*, 2022] Abhinav Joshi, Ashwani Bhat, Pradeep S, Priya Gole, Shashwat Gupta, Shreyansh Agarwal, and Ashutosh Modi. CISLR: Corpus for Indian Sign Language recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 10357–10366, 2022.

[Kumar *et al.*, 2021] Parteek Kumar, Sanmeet Kaur, et al. Indian sign language generation system. *Computer*, 54(3):37–46, 2021.

[Patra *et al.*, 2025a] Suvajit Patra, Arkadip Maitra, and Soumitra Samanta. A continuous indian sign language production system from isolated signs. In *Proceedings of the International Conference on Smart Computing and Communications*, pages 1–6, 2025.

[Patra *et al.*, 2025b] Suvajit Patra, Arkadip Maitra, Megha Tiwari, K Kumaran, Swathy Prabhu, Swami Punyeshwarananda, and Soumitra Samanta. Hierarchical windowed graph attention transformer encoder and a large scale dataset for indian sign language recognition. *Pattern Analysis and Applications*, 28(3):148, 2025.

[Rastgoo *et al.*, 2024] Razieh Rastgoo, Kourosh Kiani, Sergio Escalera, Vassilis Athitsos, and Mohammad Sabokrou. A survey on recent advances in sign language production. *Expert Systems with Applications*, 243:122846, 2024.

[Razavi *et al.*, 2019] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 32, 2019.

[Saunders *et al.*, 2020] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Progressive transformers for end-to-end sign language production. In *Proceedings of the European Conference on Computer Vision*, pages 687–705, 2020.

[Saunders *et al.*, 2022] Ben Saunders, Necati Cihan Camgoz, and Richard Bowden. Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 5141–5151, 2022.

[Shan *et al.*, 2023] Wenkang Shan, Zhenhua Liu, Xinfeng Zhang, Zhao Wang, Kai Han, Shanshe Wang, Siwei Ma, and Wen Gao. Diffusion-based 3d human pose estimation with multi-hypothesis aggregation. In *Proceedings of the International Conference on Computer Vision*, pages 14761–14771, 2023.

[Shi *et al.*, 2024] Tongkai Shi, Lianyu Hu, Fanhua Shang, Jichao Feng, Peidong Liu, and Wei Feng. Pose-guided fine-grained sign language video generation. In *Proceedings of the European Conference on Computer Vision*, pages 392–409, 2024.

[Sridhar *et al.*, 2020] Advaith Sridhar, Rohith Gandhi Ganesan, Pratyush Kumar, and Mitesh Khapra. Include: A large scale dataset for indian sign language recognition. In *Proceedings of the ACM International Conference on Multimedia*, pages 1366–1375, 2020.

[Stoll *et al.*, 2018] Stephanie Stoll, Necati Cihan Camgöz, Simon Hadfield, and Richard Bowden. Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the British Machine Vision Conference*, 2018.

[Sugandhi *et al.*, 2020] Sugandhi, Parteek Kumar, and Sanmeet Kaur. Sign language generation system based on indian sign language grammar. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 19(4):1–26, 2020.

[Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the conference on computer vision and pattern recognition*, pages 5693–5703, 2019.

[Tang *et al.*, 2022] Shengeng Tang, Richang Hong, Dan Guo, and Meng Wang. Gloss semantic-enhanced network with online back-translation for sign language production. In *Proceedings of the ACM International Conference on Multimedia*, pages 5630–5638, 2022.

[Tang *et al.*, 2025a] Shengeng Tang, Jiayi He, Lechao Cheng, Jingjing Wu, Dan Guo, and Richang Hong. Discrete to continuous: Generating smooth transition poses from sign language observations. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 3481–3491, 2025.

[Tang *et al.*, 2025b] Shengeng Tang, Jiayi He, Dan Guo, Yanyan Wei, Feng Li, and Richang Hong. Sign-idd: Iconicity disentangled diffusion for sign language production. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7266–7274, 2025.

[Tang *et al.*, 2025c] Shengeng Tang, Feng Xue, Jingjing Wu, Shuo Wang, and Richang Hong. Gloss-driven conditional diffusion models for sign language production. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(4):1–17, 2025.

[Van Den Oord *et al.*, 2017] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

[Viegas *et al.*, 2023] Carla Viegas, Mert Inan, Lorna Quandt, and Malihe Alikhani. Including facial expressions in contextual embeddings for sign language generation. In *Proceedings of the Joint Conference on Lexical and Computational Semantics*, pages 1–10, 2023.

[Walsh *et al.*, 2024] Harry Walsh, Abolfazl Ravanshad, Mariam Rahmani, and Richard Bowden. A data-driven representation for sign language production. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*, pages 1–10, 2024.

[Walsh *et al.*, 2025] Harry Walsh, Ed Fish, Ozge Mercanoglu Sincan, Mohamed Ilyes Lakhal, Richard Bowden, Neil Fox, Bencie Woll, Kepeng Wu, Zecheng Li, Weichao Zhao, et al. Slrtp2025 sign language production challenge: Methodology, results and future work. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 4109–4119, 2025.

[Wang *et al.*, 2025] Xu Wang, Shengeng Tang, Peipei Song, Shuo Wang, Dan Guo, and Richang Hong. Linguistics-vision monotonic consistent network for sign language production. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2025.

[Williams *et al.*, 2020] Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33:4524–4535, 2020.

[Xie *et al.*, 2024] Pan Xie, Qipeng Zhang, Peng Taiying, Hao Tang, Yao Du, and Zexian Li. G2p-ddm: Generating sign pose sequence from gloss sequence with discrete diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6234–6242, 2024.

[Yin *et al.*, 2024] Aoxiong Yin, Haoyuan Li, Kai Shen, Siliang Tang, and Yueting Zhuang. T2s-gpt: Dynamic vector quantization for autoregressive sign language production from text. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3345–3356, 2024.

[Zelinka and Kanis, 2020] Jan Zelinka and Jakub Kanis. Neural sign language synthesis: Words are our glosses. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 3395–3403, 2020.

[Zhang *et al.*, 2019] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

[Zhang *et al.*, 2025] Shijun Zhang, Hongkai Zhao, Yimin Zhong, and Haomin Zhou. Why shallow networks struggle to approximate and learn high frequencies. *Information and Inference: A Journal of the IMA*, 14(3):iaaf022, 2025.

[Zhou *et al.*, 2021] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li. Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 1316–1325, 2021.

[Zuo *et al.*, 2024] Ronglai Zuo, Fangyun Wei, Zenggui Chen, Brian Mak, Jiaolong Yang, and Xin Tong. A simple baseline for spoken language to sign language translation with 3d avatars. In *Proceedings of the European Conference on Computer Vision*, pages 36–54, 2024.

[Zwitserlood *et al.*, 2004] Inge Zwitserlood, Margriet Verlinden, Johan Ros, Sanny Van Der Schoot, and T Netherlands. Synthetic signing for the deaf: Esign. In *Proceedings of the Conference and Workshop on Assistive Technologies for Vision and Hearing Impairment*, 2004.